DOCUMENT RESUME

ED 447 167 TM 032 009

AUTHOR De Ayala, R. J.; Plake, Barbara S.; Impara, James C.; Kozmicky, Michelle

TITLE The Effect of Omitted Responses on Ability Estimation in IRT.

PUB DATE 2000-00-00

NOTE 48p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Ability; *Estimation (Mathematics); *Item Response Theory; Maximum Likelihood Statistics; Responses; Simulation; *Test Items

IDENTIFIERS Expected A Posteriori Tests; *Omitted Responses; Three Parameter Model

ABSTRACT

This study investigated the effect on examinees' ability estimate under item response theory (IRT) when they are presented an item, have ample time to answer the item, but decide not to respond to the item. Simulation data were modeled on an empirical data set of 25,546 examinees that was calibrated using the 3-parameter logistic model. The study was conducted in three phases. Phase 1 was an exploratory study comparing the various estimation methods under different conditions. Methods compared were the Biweight, expected a posteriori (EAP), and maximum likelihood estimation (MLE) methods. Phase 2, based on Phase 1 results, examined a modified EAP approach. The third phase examined the use of R. Bock's 1972 nominal response (NR) model for handling omission data. Results seem to indicate that omits should not be treated as incorrect. It also appears that ignoring omits can have a greater impact on thetas using certain estimation approaches (e.g., MLE) than with others. To the extent that a model with a pseudo-guessing parameter more accurately describes the data than a model without one, then the use of the NR model may not produce results comparable to those seen here. (Contains 5 tables, 10 figures, and 14 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

The Effect of Omitted Responses on Ability Estimation in IRT

R.J. De Ayala, Barbara S. Plake, James C. Impara, and Michelle Kozmicky
University of Nebraska-Lincoln

ED 447 167

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

R. DeAyala

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM032009

Paper presented at the Annual Meeting of the American Educational Research Association, April 24-28, New Orleans, LA

BEST COPY AVAILABLE

The Effect of Omitted Responses on Ability Estimation in IRT

For a number of reasons an examinee's response vector may not contain responses to each item. For example, the items not presented in an adaptive test or the non-common items in a common-item equating design will only have responses for a subset of examinees. Both of these examples share the characteristic that the test administration involves a decision to not present certain items to all examinees. Using Little and Rubin's (1987) terminology these nonresponses represent conditions in which the missingness process may be ignored for purposes of ability estimation (Mislevy & Wu, 1988; Mislevy & Wu, 1996). In contrast, "not-reached" items are items that an examinee is unable to consider answering because of insufficient time. These not-reached items can be identified as collectively occurring at the end of an exam (this assumes the examinee responds to the test items in serial order). Lord (1980) stated that in practice these not-reached item may be ignored for ability estimation because they contain no readily quantifiable information about the examinee's ability. Augmenting this perspective, Mislevy and Wu (1996) outlined the conditions in which not-reached items may represent ignorable missing data. Another source of missing data occurs because examinees have the capability of choosing not to respond to certain questions on an examination. These (intentionally) omitted responses represent nonignorable missing data (Lord, 1980; Mislevy & Wu, 1988; Mislevy and Wu, 1996). This study investigated the effect on an examinee's ability estimate when he or she is presented an item, has ample time to answer the item, but decides to not respond to the item.

It is reasonable to believe that, in general, an examinee who omits responding to an item does so because the examinee believes that he or she does not know the answer to the question. A highly proficient individual, by virtue of his or her ability, may be more likely to realize that he or she does not know the answer to an item better than a less proficient examinee. Therefore, the highly proficient examinee will have a greater tendency to omit items that he or she does not know the answers to than does a less proficient examinee. In addition, the highly proficient examinee may tend to omit responses at a lower rate than does a less proficient examinee. As a result, the highly proficient examinee's response string will tend to contain more correct responses than if the examinee had responded to the

omitted items and the number of omissions will be less than that of a less proficient examinee. Conversely, a less proficient examinee may not be able to make the distinction that he or she knows the answer to a question as well as a highly proficient examinee and as a consequence will tend to omit items that the examinee may have correctly responded to if he or she had answered the question (cf. Wainer & Thissen, 1994). Clearly, in the context of ability estimation omitted responses are not ignorable because the act of omission is related, in part, to the examinee's ability. Lord (1980) has argued that omitted responses may not be ignored because an examinee that understands ability estimation in the context of item response theory (IRT) could obtain as high an ability estimate as he or she wished by simply answering only those items he or she has confidence in correctly answering. This idea has found some support in Wang, Wainer, and Thissen's (1995) study on examinee item choice.

There are a number of different ability estimation approaches in IRT with different advantages and disadvantages. It might be expected that the effect of omitted responses on ability estimation may vary as a function of estimation approach. For instance, if a Bayesian-based method is used, then the regression toward the mean phenomenon inherent in a Bayesian approach might be expected to compensate to some extent for the potential underestimation of less proficient examinees and the overestimation of highly proficient examinees. In contrast, a maximum likelihood-based approach might be expected to show the aforementioned biases. A procedure proposed by Mislevy and Bock (1982), biweight ability estimation, was developed to provide robust ability estimation using maximum likelihood. With this method the likelihood is modified to weight items closer to the examinee's proficiency more than those further away. Weighting the items appropriately may provide a means of compensating for the expected biases and result in a more accurate ability estimate than would be obtained using a nonweighted maximum likelihood approach. An alternative approach to dealing with missing data is based on Lord (1974). This method involves the assignment of a fractionally correct value equal to the reciprocal of the number item alternatives (i.e., the random guessing value) to the omitted item(s). This latter method assumes that examinees omit items if their chances of correctly responding would have been equal to random guessing. In addition, this approach assumes that both highly and less proficient examinees can be treated the same. However, this

assumption may not be tenable because Stocking, Eignor, and Cook (1988) have shown that the rates of omission vary as a function of ability. Moreover, Mislevy and Wu (1988) have stated that the tendency to omit can be associated with personality characteristics, demographic variables, as well as ability level. Therefore, differential omission rates may not be compensated for using Lord's approach for proficiency estimation.

Method

Ability Estimation Methods

Ability estimation has typically used either maximum likelihood estimation (MLE) or a Bayesian approach such as maximum a posteriori (MAP or Bayes Model Estimate) or expected a posteriori (EAP or Bayes Mean Estimate). The former two algorithms are iterative techniques, while EAP is non-iterative and is based on numerical quadrature methods. Unlike MLE ability estimates, EAP ability estimates may be obtained for all response patterns, including zero and perfect score patterns. While MAP proficiency estimates also exist for all response patterns, they suffer from greater regression towards the mean than do the EAP estimates (Bock & Mislevy, 1982; Mislevy & Bock, 1990). The EAP estimate (Bock & Mislevy, 1982) of an examinee's proficiency, θ, after n items have been administered is given by

$$(\hat{\theta})_n = \frac{\sum_{k=1}^{q} X_k L_n(X_k) A(X_k)}{\sum_{k=1}^{q} L_n(X_k) A(X_k)} \qquad (1)$$

and its posterior standard deviations is

$$\mathrm{PSD}(\hat{\theta}) = \sqrt{\frac{\sum_{k=1}^{q} (X_k - \hat{\theta}_n)^2 L_n(X_k) A(X_k)}{\sum_{k=1}^{q} L_n(X_k) A(X_k)}} \qquad (2)$$

where X_k is one of q quadrature points, $A(X_k)$ is the corresponding quadrature weight, and $L_n(X_k)$ is the likelihood function of X_k given the response pattern $[x_1, x_2, ..., x_n]$. For example, if the probability of a correct response by an individual with proficiency θ to a dichotomously scored item i with location b_i,

discrimination a_i, and pseudo-guessing parameter c_i is given by the three-parameter logistic (3PL) model

$$p(x_i = 1|\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta - b_i)}} \tag{3}$$

then the likelihood of θ given the response pattern $\{x_1, x_2, ..., x_n\}$ is

$$L_n(\theta) = \prod_{i=1}^{n} p(x_i = 1|\theta)^{x_i} (1 - p(x_i = 1|\theta))^{(1-x_i)}. \tag{4}$$

MLE uses a gradient approach for determining the location of the maximum of (4) (i.e., the value that maximizes the likelihood). This location is taken as the examinee's $\hat{\theta}$. In practice the natural log of (4) is typically used. Given some estimate of an examinee's θ, $\hat{\theta}^t$, the estimate is refined by examining the average rate of change of the function with respect to a particular point. Technically, this refinement takes the form of a ratio (Δ) of the first derivative to the second derivative of the log likelihood function (Lord, 1980)

$$\frac{\partial \ln L}{\partial \theta} = \frac{\sum_{i=1}^{n} a_i(x_i - p_i)(p_i - c_i)}{p_i(1 - c_i)} \tag{5}$$

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{\sum_{i=1}^{n} a_i^2(x_i c_i - p_i^2)(p_i - c_i)(1 - p_i)}{p_i^2(1 - c_i)^2} \tag{6}$$

where p_i is defined by (3) given the appropriate item parameters and current $\hat{\theta}$. Therefore, the refinement of $\hat{\theta}$ at the t+1 iteration is given by

$$\hat{\theta}^{t+1} = \hat{\theta}^t - \Delta = \hat{\theta}^t - \frac{\frac{\partial \ln L}{\partial \theta}}{\frac{\partial^2 \ln L}{\partial \theta^2}} \tag{7}$$

Iterations continue until $\hat{\theta}^{t+1}$ is considered to be equivalent to $\hat{\theta}^t$ to some degree of accuracy. At this point the examinee's $\hat{\theta}$ is taken to be $\hat{\theta}^{t+1}$. The estimate of the standard error of estimate for $\hat{\theta}$ is

$$SEE(\hat{\theta}) = \sqrt{\frac{1}{\sum_{i=1}^{n} \frac{[a_i(1 - p_i)(p_i - c_i)/(1 - c_i)]^2}{p_i(1-p_i)}}} \tag{8}$$

and p_i is from the final iteration.

Mislevy and Bock (1982) introduced a modification of MLE to reduce its sensitivity to responses that are inconsistent with an IRT model (e.g., (3)). An example of such a "response disturbance" would be an incorrect response to an "easy" item by a high-ability examinee. Their modification involved the application of Tukey's biweight to the estimation of θ. The biweight is primarily inversely related to the distance between an item's location and the examinee's biweight $\hat{\theta}$. The closer the item is to the examinee, the greater the weight given to the item, and the further away the item is from the examinee, the less the weight that is given to the item. In short, unexpected responses are given less weight than responses that are consistent with the model. Estimation proceeds as in (7) except that the ratio of derivatives is modified to include "item weights" (W_i) that are iteration specific

$$\Delta = \frac{\sum W_i^t a_i(x_i - p_i^t)}{-\sum W_i^t a_i^2 p_i^t(1 - p_i^t)} \qquad (9)$$

where if $|U_i^t| \leq 1$, then $W_i^t = [1 - (U_i^t)^2]^2$, otherwise $W_i^t = 0$, $U_i^t = a_i(b_i - \hat{\theta}^t)/C$, and p_i is defined by (3) given the appropriate item parameters and current $\hat{\theta}$. If $(a_i(b_i - \hat{\theta}^t)/C) > 1$, then the "item weight" is zero and the item is effectively "removed" or "trimmed." Therefore, there is an inverse relationship between C and the amount of trimming to be conducted. C is an arbitrary constant that specifies, as a function of the logit, the amount of trimming to be done. As was the case with MLE, iterations continue until $\hat{\theta}^{t+1}$ is considered to be equivalent to $\hat{\theta}^t$ to some degree of accuracy. This $\hat{\theta}^{t+1}$ is taken as the examinee's $\hat{\theta}$. The estimate of the standard error of estimate for $\hat{\theta}$ is

$$SEE(\hat{\theta}) = \sqrt{-\sum W_i a_i^2 p_i(1 - p_i)} \qquad (10)$$

where W_i and p_i are from the final iteration.

Data Generation:

The simulation data were modeled on a empirical data set. This empirical data set consisted of 24,546 examinees and had been calibrated using the 3PL model. Of these examinees, 6515 examinees had response vectors that contained a combination of correct, incorrect, and omitted responses to 39 items. The average number of correct responses for these 6515 examinees was 19.184 (Median=18) with a standard deviation of 8.089 (minimum score=1, maximum score=38, skew=0.300). For these latter examinees the average number of items omitted was 2.224 (Median=1) with a standard deviation of

2.851 (minimum number omitted=1, maximum number omitted=35, skew=4.251). For the 6515 examinees 96% omitted 8 items or less. Because an examinee may omit an item as a function of many different factors (e.g., knowledge of the answer, self-confidence, risk-aversion, test-wiseness, metacognitive factors, etc.) and there were no explicit measures of these factors it was decided to not use a parametric approach for modeling the empirical data. Because the omission pattern across ability differed for persons who responded correctly versus incorrectly to an item, a pair of contingency tables was created for each item using the 6515 examinees that had response vectors containing correct, incorrect, and omitted responses. Each contingency table consisted of a two-level response type variable versus an ability measure variable. For one table the response type variable consisted of response omission or responding incorrectly to the item, whereas for the other table for that item the response type variable consisted of omitting a response or correctly responding to the item. The ability measure variable consisted of ten 4-item fractiles of the number correct score (0-3, 4-7, etc.). By using ten 4-item fractiles in lieu of deciles it was felt that we would avoid having some fractiles that consisted of a relatively large range of number correct scores and others that consisted of 1 or 2 number correct scores. Based on these tables the proportion of individuals omitting a response to an item conditional on the fractile were calculated.

The simulated data were generated on the basis of (3) and the item parameter estimates of the empirical data were treated as known. For each 0.1 of logit from -2.0 to 2.0 (inclusive) 1000 θs were generated for a total of 41,000 simulees. For each simulee the probability of a correct response was calculated according to (3) and compared to a uniform random number [0,1]. If the random number was less than or equal to probability of a correct response, then the response was coded as '1' for correct, '0' otherwise. To generate the omission data, the number correct score for each simulee was determined and the simulee assigned to one of the ten fractiles. For each item the correctness of the simulee's response was used to determine which of the two contingency tables for the item should be used. Based on the simulee's fractile assignment the appropriate relative frequency of omission was compared to a uniform random number [0,1]. If the uniform random number was less than or equal to the relative frequency for omission, conditional on the simulee's fractile, then the response was changed to be an omission,

otherwise the simulee's response to the item was not changed. For instance, for an item the relative frequency of omission for an examinee in the third fractile might be 0.42 if the simulee responded incorrectly to the item and 0.11 if the simulee responded correctly. If on the basis of the data generation a simulee responded responded incorrectly to the item, then a uniform random number would be generated and compared to 0.42. If this random number was, for example, 0.3, then the simulee's incorrect response to this item would be changed to reflect that it had been omitted. This process was repeated for each of the 39 items and for all simulees. Therefore, each simulee had a response vector of correct and incorrect responses (a.k.a., the complete vector) and a response vector of correct, incorrect and omitted responses (a.k.a., the omission vector).

The study was conducted in three phases. Phase 1 was an exploratory study comparing the various estimation methods under different conditions. Phase 2 was based on Phase 1 results and examined a modified EAP approach. The third phase examined the use of Bock's (1972) nominal response (NR) model for handling omission data.

Phase 1

Factors:

The Biweight, EAP, and MLE estimation methods were investigated. For the Biweight method 5 different levels of trimming were examined (C=2, C=4, C=6, C=8, C=10), for the EAP approach two different levels of quadrature points (10 and 20 points), and for MLE the omitted responses were replaced with the reciprocals 4 and 7 (7 was approximately equal to the reciprocal of the median *c* value; this factor was called Nalt for number of alternatives). In addition, for the MLE method omitted responses were treated as Incorrect as well as Ignored for ability estimation. Each simulee's ability was estimated using each estimation method. For each method each simulee had two $\hat{\theta}$s: one $\hat{\theta}$ based on the simulee's complete vector ($\hat{\theta}_C$) and the other $\hat{\theta}$ using the simulee's omission vector ($\hat{\theta}_O$). All methods used (3).

Each level of the ability estimation methods was crossed by the number of items omitted in the response vector (Nomitted). Nomitted consisted of four levels: 2, 4, 6, and 8 omitted responses (for the

simulated data the cumulative percent for omitting 8 items or less was 99.5%). These four levels of Nomitted, 2, 4, 6, and 8, represent 5.1%, 10.3%, 15.4%, 20.5% of the test length, respectively.

Analysis:

Descriptive statistics were calculated on the item parameters and ability estimates. Fidelity coefficients were obtained. Each ability estimate's Root Mean Square Error (RMSE) and Bias were calculated. RMSE was calculated according to:

$$RMSE(\theta_k) = \sqrt{\frac{\sum (\hat{\theta} - \theta_k)^2}{n_k}} \qquad (11)$$

$$Bias(\theta_k) = \frac{\sum (\hat{\theta} - \theta_k)}{n_k} \qquad (12)$$

where $\hat{\theta}$: proficiency estimate based on one of the estimation methods using either the complete or omission vectors

θ_k: simulee's proficiency at logit k (-2.0, -1.9, -1.8, ..., 2.0)

n: the number of simulees at logit k

RMSE and Bias were calculated separately for the complete vectors and omission vectors. Because RMSEs for the complete vectors represented how well the simulees could be estimated on the basis of complete response data, the RMSEs for the omission vectors were compared to the corresponding RMSEs for the complete vectors; this was also true for Bias. These differences between the RMSE for the omission and complete vectors as well as for Bias were examined graphically for each condition. All statistics were calculated using convergent cases.

Programs:

To perform the ability estimation Biweight, EAP, and MLE programs were written. A program to calculate RMSE and Bias was also written.

Phase 2

Based on the results of Phase 1 another condition was implemented. The same analysis measures and data used in Phase 1 were used for Phase 2. The results for MLE using 4 and 7 as the number of alternatives indicated that using 2 as the number of alternatives may be productive. EAP was selected as the ability estimation method for Phase 2 because it is a noniterative method for which finite

ability estimates are always available and because, on average, its performance was better than MLE. Because the comparison of EAP results using 10 quadrature points were very similar to those using 20 points and both MULTILOG (Thissen, 1991) and BILOG (Mislevy & Bock, 1990) use 10 quadrature points as default for EAP estimation, the number of quadrature points used in Phase 2 was 10.

Phase 3

To explore the use of Bock's (1972) NR model for handling omits the simulated data set of 41,000 simulees was calibrated for the NR and two-parameter logistic (2PL) models; the 2PL is mathematically equivalent to the 3PL , but with *c* set to 0.0. For the 2PL model the complete vectors were used for item calibration. For the NR model omits were coded 1, while incorrect and correct responses were coded 2 and 3, respectively. Using the appropriate item parameter estimates the simulees' θs were estimated using MAP for both the 2PL and NR models. In addition, the item parameters used for data generation were converted to their corresponding contrast coefficients and used for estimating the simulees' θs according to the 3PL model (MAP estimation). The same analysis measures used in Phase 1 were used for Phase 3.

Results

Item pool:

Table 1 contains descriptive statistics for the item pool used. As can be seen the item locations were distributed between -2.26 and 1.3 and centered at -0.5547 with an average item discrimination of 0.8866. The correlations between the number of times an item was omitted and item discrimination, location, and intercept were 0.0338, -0.3291, and 0.3509, respectively. The maximum test information was approximately 5.19 and was located at -0.2185.

Insert Table 1 about here

The four levels of Nomitted consisted of 9713 simulees that omitted two items, 6948 that omitted four items, 2229 simulees that omitted six items, and 431 that omitted eight items. For these levels the average trait values were $\bar{\theta}_2$ = 0.3604 (SD=1.1332), $\bar{\theta}_4$ = -0.3335 (SD=1.0830), $\bar{\theta}_6$ = -0.8139 (SD=0.9029), and $\bar{\theta}_8$ = -1.0694 (SD=0.7721).

Phase 1

Table 2 shows the fidelity coefficients as well as the intercorrelation between the Biweight ability estimates based on the complete vectors and the omission vectors; $\bar{\hat{\theta}}_C$ and $\bar{\hat{\theta}}_O$ represent the mean estimates using the complete and omission vectors, respectively. For all levels of the trimming factor the $r_{\theta\hat{\theta}_C}$s were greater than the $r_{\theta\hat{\theta}_O}$s for corresponding Nomitted levels. As would be expected, as the number of omits increased the fidelity coefficients decreased for a given trim level. These decreases were similar across the trim factor levels; there was approximately a difference of 0.09 between the largest and smallest fidelity coefficients. Although the $r_{\theta\hat{\theta}_O}$s were greater for the higher trimming levels (i.e., C=2) than for lower trim levels (i.e., C=10), the differences between corresponding $r_{\theta\hat{\theta}_O}$s for similar Nomitted levels were slight. The $r_{\hat{\theta}_C\hat{\theta}_O}$s tended to increase as less trimming was used on the ability estimates for corresponding levels of Nomitted. In general, for a given trim level the $r_{\hat{\theta}_C\hat{\theta}_O}$s decreased with increasing level of Nomitted, although they were still greater than 0.94.

Insert Table 2 about here

Comparison of the $r_{\theta\hat{\theta}_O}$s across the levels of the number of quadrature points factor for corresponding Nomitted levels showed that these correlations varied by less than 0.0007 (Table 3). The fidelity coefficients involving the complete vectors ($r_{\theta\hat{\theta}_C}$) for 10 quadrature points differed from the $r_{\theta\hat{\theta}_C}$s for 20 quadrature points by 0.0003 or less. The only exception to this occured for the Nomitted=6 level in which the fidelity coefficients differed by 0.001. As was the case for the Biweight estimation, the $r_{\theta\hat{\theta}_C}$s were greater than the $r_{\theta\hat{\theta}_O}$s for corresponding Nomitted levels. The correlation between the $\hat{\theta}_C$ and $\hat{\theta}_O$ showed the same pattern as was seen with the Biweight estimates, although the EAP correlations were greater than 0.96.

Insert Table 3 about here

Of the three estimation methods, MLE showed the lowest fidelity coefficients for both $\hat{\theta}_C$ and $\hat{\theta}_O$ for corresponding levels of omission (Table 4). As was the case with the Biweight and EAP $\hat{\theta}$s, as the

number of omits increased the fidelity coefficients decreased, although the difference between the largest and smallest value for given Nalt level was larger than that seen with the other ability estimation methods (this difference was as large as 0.1572 in one condition). For a given Nomitted level, the lowest $r_{\theta\hat{\theta}_C}$s were seen when omitted responses were ignored when estimating the simulee's ability. This was also the condition in which the largest number of nonconvergent cases were observed, although proportionally the nonconvergent cases represented less than 1% of the cases estimated. This finding was not expected. It was anticipated that the number of nonconvergent cases would increase as the number of omissions increased and the number of these cases would be larger than what was observed.

Insert Table 4 about here

The accuracy of estimation was studied graphically. For all figures only data points based on 10 or more cases were plotted. Figure 1 contains RMSE as a function of θ. The bold dashed line represents RMSE($\hat{\theta}_C$) while the remaining nonbold lines represent the difference between RMSE based on the omission vector (RMSE($\hat{\theta}_O$)) and RMSE based on the complete vector (RMSE($\hat{\theta}_C$)) for a given level of Nomitted; this is true for all the RMSE plots discussed. Values above the baseline indicate that RMSE($\hat{\theta}_O$) was greater than RMSE($\hat{\theta}_C$). The patterm of RMSE($\hat{\theta}_C$) was what would be expected given the unimodal test information function. In general, the accuracy of the Biweight ("heavily trimmed", C=2) $\hat{\theta}_O$s was slightly less than that based on the complete data across the proficiency continuum (Figure 1a). In general, increasing levels of omission led to slightly larger discrepancies between RMSE($\hat{\theta}_C$) and RMSE($\hat{\theta}_O$), however, this effect of omission at the lower end of the continuum is not very large. The most erratic pattern observed corresponded to cases with 8 omitted responses and even for these cases this occurred over a limited θ range. Because the maximum location parameter was 1.29, the patterns displayed above this location may be somewhat idiosyncratic.

Insert Figure 1 about here

Figure 1b contains the the Biweight C=6 RMSE results (the C=4 condition falls predicably between this figure and Figure 1a). This condition represents less trimming than that in Figure 1a. However, except for slight increases in RMSE($\hat{\theta}_O$) for the 8 omitted responses, there appeared to be little effect due to reducing the amount of trimming for simulees located below approximately 0.5. The pattern continued to be exhibited when trimming was further reduced (e.g., Figure 1c).

The corresponding Biweight Bias plots are presented in Figure 2. For the Bias figures the bold dashed line represents Bias($\hat{\theta}_C$) while the remaining nonbold lines represent the difference between Bias based on the omission vector (Bias($\hat{\theta}_O$)) and Bias based on the complete vector (Bias($\hat{\theta}_C$)) for a given level of Nomitted; this is true for all the Bias plots discussed. As can be seen, the Biweight $\hat{\theta}_C$s tended to underestimate low θs and overestimate at the upper proficiency levels. At θs less than 1.0 trimming did not eliminate this pattern of under- and overestimation. In general, for the C=2 condition (Figure 2a) increasing omission levels led to increasing levels of bias in $\hat{\theta}_O$. This pattern can also be observed in the C=6 (Figure 2b) and C=10 (Figure 2c) conditions. All figures showed a pattern of increasing Bias($\hat{\theta}_O$) as Nomitted increased. The negative differences observed indicate a situation in which there was less bias in $\hat{\theta}_O$ than in $\hat{\theta}_C$, although, as stated above, these $\hat{\theta}_O$ may be somewhat less stable than those below θ = 1.29.

Insert Figure 2 about here

Figures 3 and 4 contain the EAP RMSE and Bias results for the 10 quadrature point condition (the results for the 20 quadrature point level are very similar to these figures). A comparison of EAP RMSE($\hat{\theta}_C$) with the Biweight RMSE($\hat{\theta}_C$) showed that EAP RMSE($\hat{\theta}_C$)s were approximately 0.03 less than the Biweight $\hat{\theta}_C$s towards the ends of the continuum, whereas the Biweight RMSE($\hat{\theta}_C$) were 0.018 less than the EAP RMSE($\hat{\theta}_C$) in the center of the scale. As was the case with Biweight estimation, increasing levels of omission led to slightly larger RMSE($\hat{\theta}_O$), however, this effect of omission at the lower end of the continuum was not very large and the most erratic pattern observed corresponded, as above, to the 8 omitted responses cases. In general, the fewer the number of omissions the more similar RMSE($\hat{\theta}_C$) and RMSE($\hat{\theta}_O$) were.

Insert Figure 3 about here

As one would expect from EAP, the $\hat{\theta}_C$s tended to underestimate low θs and overestimate at the upper proficiency levels (Figure 4). Unlike the Biweight condition, EAP Bias($\hat{\theta}_O$) appeared to increase around θ = 0.0 for all levels of Nomitted, although the pattern of increasing Bias($\hat{\theta}_O$) as Nomitted increased was still evident above approximately θ = -1.0. Bias($\hat{\theta}_C$) and Bias($\hat{\theta}_O$) were virtually identical at the lowest end of the θ continuum.

Insert Figure 4 about here

Figure 5 contains the RMSE plots for MLE ability estimation. Comparing MLE RMSE($\hat{\theta}_C$) with those of Biweight and EAP showed that MLE was not as accurate, on average, as Biweight and EAP for θ > -1.40. The pattern of increasing RMSE($\hat{\theta}_O$) as a function of increasing Nomitted presented above was also observed with MLE. Although the Figures 5a, 5b, and 5d show that for the Nomitted=2 condition RMSE($\hat{\theta}_O$) was less than RMSE($\hat{\theta}_C$), the difference was relatively small and may be attributed to random sampling fluctuations; t-tests on the ln(RMSE) showed that there was no statistically significant (α = 0.05) differences between RMSE($\hat{\theta}_O$) and RMSE($\hat{\theta}_C$) for Nalt levels of 4 and 7, and Treating Omits as Incorrect. The effect of the number of omits on the accuracy of $\hat{\theta}_O$ was more pronounced with MLE than was observed with either Biweight or EAP. It appeared that MLE ability estimation was not very affected by two or four omissions (10.3% or fewer omits), but omitting more than 4 items had a marked increase in RMSE($\hat{\theta}_O$) (Figures 5a and 5b). Comparing Figures 5a and 5b it can be seen that increasing the number of alternatives from 4 to 7 led to increases in RMSE($\hat{\theta}_O$) for Nomitted=4, 6, and 8 conditions. Ignoring omits in estimating proficiency decreased the accuracy of $\hat{\theta}_O$ above θ = -0.5, but below this point the differences between RMSE($\hat{\theta}_O$) and RMSE($\hat{\theta}_C$) were similar to those observed when the number of alternatives was 4 (Figure 5c). Figure 5d showed that treating the omitted responses as

incorrect led to the largest discrepancies between RMSE($\hat{\theta}_O$) and RMSE($\hat{\theta}_C$) of all conditions investigated.

Insert Figure 5 about here

The corresponding Bias plots for the four MLE approaches are presented in Figure 6. Unlike Biweight and EAP, there was, on a average, a relatively consistent positive Bias($\hat{\theta}_C$) throughout the proficiency continuum for MLE. In contrast to Biweight and EAP and except for the Ignoring Omits condition, the relationship of Bias($\hat{\theta}_O$) and Bias($\hat{\theta}_C$) was different than previously observed. Specifically, for omission vectors with two omits Lord's approach led to less positively biased $\hat{\theta}_O$ than would normally be observed with complete vectors. This was also true for treating omits as incorrect. However, inspection of the Bias($\hat{\theta}_O$)s for the 4, 6, and 8 Nomitted levels showed increasing negatively biased $\hat{\theta}_O$s as a direct function of Nomitted. Moreover, for the 4, 6, and 8 Nomitted levels the largest negatively biased $\hat{\theta}_O$s were found when omits were treated as incorrect and the smallest when specifying 4 alternatives. Ignoring omits in estimating $\hat{\theta}$ (Figure 6c) showed greater bias in $\hat{\theta}_O$ than that observed in $\hat{\theta}_C$. For the Nomited=2 and 4 levels, $\hat{\theta}_O$ was more positively biased than $\hat{\theta}_C$ throughout the θ range. With Nomitted=6, there was a negative bias in $\hat{\theta}_O$ at lower θs (Bias($\hat{\theta}_C$) was positive in this range) that became a positive bias as θ increased (Bias($\hat{\theta}_C$) became negative as θ increased). For Nomitted=8, Bias($\hat{\theta}_O$) and Bias($\hat{\theta}_C$) were, in general, negatively biased throughout the proficiency continuum, although $\hat{\theta}_O$ showed greater negative bias than $\hat{\theta}_C$.

Insert Figure 6 about here

Phase 2

Given that the above RMSE and Bias results for Nalt=4 were better than that for Nalt=7 it was hypothesized that specifying that the number of alternatives as 2 might further reduce RMSE($\hat{\theta}_O$) and Bias($\hat{\theta}_O$). Although EAP showed RMSE($\hat{\theta}_O$)s that were less than that of MLE, MLE showed less bias at

the ends of the proficiency continuum than did EAP. To determine whether EAP or MLE was, overall, performing better, the variance error of estimate was calculated from RMSE and Bias by:

$$VEE(\theta) = RMSE(\theta)^2 - Bias(\theta)^2 \qquad (CC)$$

The square root of VEE is presented in Figure 7. As can be seen EAP performed better than MLE across the θ continuum.

Insert Figure 7 about here

Specifying that the number of alternatives was 2 and using 10 quadrature points, EAP $\hat{\theta}$s were obtained for each level of Nomitted. The corresponding RMSE($\hat{\theta}_O$)s and Bias($\hat{\theta}_O$)s are presented in Figure 8. Figure 8a shows that the agreement between RMSE($\hat{\theta}_O$) and RMSE($\hat{\theta}_C$) increased relative to what had been observed in Figure 3 for all Nomitted levels. Similarly, the discrepancy between Bias($\hat{\theta}_O$) and Bias($\hat{\theta}_C$) was also reduced (cf. Figure 4).

Insert Figure 8 about here

Phase 3

Table 5 shows the fidelity coefficients for the 2PL, 3PL, and NR models. For all levels of Nomitted the fidelity coefficients for the NR model were less than those of the 2PL and 3PL models. On average, the differences between these coeffecients for dichotomous models and NR models were 0.0173 or less. Given that the NR models subsumes the 2PL model, it was not surprising to find the stronger agreement between the $\hat{\theta}$s for the 2PL and NR models than between the 3PL and NR models. The regression toward the mean expected of MAP estimates was reflected in standard deviations for the $\hat{\theta}$s that were less than of θ.

Insert Table 5 about here

Figures 9 and 10 contain the RMSE and Bias plots for the dichotomous and NR models. The bold dash line in each Figures 9a and 10a represents RMSE($\hat{\theta}_C$), while in Figures 9b and 10b it reflects

Bias($\hat{\theta}_C$). The remaining nonbold lines represent the difference between RMSE based on the omission vectors (i.e., NR model) and the RMSE based on the complete vector (i.e., either the 2PL or 3PL models) for a given level of Nomitted (Figures 9a and 10a). For the Bias figures (Figures 9b and 10b), the nonbold lines represent the difference between Bias based on the omission vectors and Bias based on the complete vectors for a given level of Nomitted. A comparison of Figures 9a and 10b showed that the 2PL model was slightly more accurate in estimation in the center of the continuum, whereas the 3PL was slightly more accurate at the ends of the continuum. The expected underestimation of high θ and overestimation of low θ that was not seen with MLE estimation (e.g., Figure 6) is clearly visible in Figures 9b and 10b. As can be seen from Figures 9a and 9b the NR model closely approximated the RMSE and Bias values for the 2PL model with slightly greater discrepancies with increased levels of Nomitted.

Insert Figure 9 about here

Figure 10a shows a greater discrepancy between the 3PL (complete data) and NR models than was seen with the 2PL. However, this discrepancy between the results of the 3PL and NR models was not as large as seen with Phase 1's Biweight, EAP, and MLE using only dichotomous models.

Insert Figure 10 about here

Discussion

The above results seem to indicate that omits should not be treated as incorrect. It also appears that ignoring omits can have a greater impact on $\hat{\theta}$s using certain estimation approaches (e.g., MLE) than with others. For this study the data were generated, in part, according to a 3PL model. The results involving a NR model as well as the 2PL, showed that they were able to approximate the accuracy of the 3PL model. It should be noted that in the comparison of the 3PL and NR models (Phase 3), the 3PL model $\hat{\theta}$s utilized the item parameters used for data generation, whereas, the NR model used item parameter estimates. Therefore, part of the discrepancy seen in the 3PL /NR RMSE as well

as 3PL/NR Bias may be, in part, attributable to this issue; this is also true the results involving the 2PL $\hat{\theta}$s. To the extent that a model with a pseudo-guessing parameter more accurately describes the data than a model without one, then the use of the NR model may not produce results comparable to those seen here. In these cases, EAP using Lords approach, but with the number of alternatives set at two, may be an estimation method that should be considered. In this case Nalt is a misnomer and should not be interpreted to indicate a two-alternative item. Instead of assuming that all examinees would answer an item (instead of omitting it) if their chances of correctly answering it were greater than 1/Nalt (i.e., the random guessing value), specifying Nalt=2 simply minimizes the magnitude of the possible discrepancy between the expected (using random guessing model) and the predicted probability of a correct response based an IRT model. As such, one is simply imputing a "response" for a binomial variable and thereby "smoothing" irregularities in the likelilhood function.

References

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Little, R., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lord, F.M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247-264.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R.J., & Bock, R.D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement,* 42, 725-737.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring: with binary logistic model.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R.J., & Wu, P. (1988). *Inferring examinee ability when some item responses are missing* (RR 88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R.J., & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (RR 96-30-ONR). Princeton, NJ: Educational Testing Service.

Stocking, M.L., Eignor, D., & Cook, L. (1988). *Factors affecting the sample invariant properties of linear and curvilinear observed and true score equating procedures* (RR-88-41). Princeton, NJ: Educational Testing Service.

Thissen, D.J. (1991). *MULTILOG* (Version 6.0). Scientific Software, Inc. Mooresville, IN.

Wainer, H., & Thissen (1994). On examinee choice in educational testing. *Review of Educational Research, 64,* 159-195.

Wang, X.B., Wainer, H., & Thissen (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education, 8,* 211-225.

Table 1: Item Pool Descriptive Statistics

	Item Parameters		
	a	*b*	*c*
Mean	0.8866	-0.5547	0.1644
Median	0.8815	-0.5350	0.1347
Standard dev	0.2621	0.8194	0.1271
Minimum	0.4168	-2.2552	0.0000
Maximum	1.5572	1.2916	0.4388

Item Parameter Intercorrelations			
	a	*b*	*c*
a			
b	0.3505		
c	0.3422	0.5918	

Table 2: Descriptive Statistics and Fidelity Coefficients-Biweight

Method	Level	Nomitted	$r_{\theta\hat{\theta}_c}$	$r_{\theta\hat{\theta}_o}$	$r_{\hat{\theta}_c\hat{\theta}_o}$	$\bar{\hat{\theta}}_c$	$\bar{\hat{\theta}}_o$	Nonconvergent[b] Compl	Omit
Biweight	C=2	2	0.9185	0.9066	0.9774	0.3822	0.2984	0	0
		4	0.9099	0.8931	0.9768	-0.2363	-0.1888	0	0
		6	0.8690	0.8457	0.9671	-0.6432	-0.5305	0	0
		8	0.8319	0.8178	0.9466	-0.9106	-0.7606	0	0
	C=4	2	a	0.9148	0.9914	a	0.3991	0	0
		4	a	0.8997	0.9871	a	-0.1570	0	0
		6	a	0.8527	0.9769	a	-0.5445	0	0
		8	a	0.8248	0.9598	a	-0.8009	0	0
	C=6	2	a	0.9144	0.9928	a	0.4297	0	0
		4	a	0.8998	0.9881	a	-0.1466	0	0
		6	a	0.8530	0.9776	a	-0.5470	0	0
		8	a	0.8247	0.9615	a	-0.8115	0	0
	C=8	2	a	0.9139	0.9930	a	0.4417	0	0
		4	a	0.8996	0.9882	a	-0.1425	0	0
		6	a	0.8529	0.9777	a	-0.5479	0	0
		8	a	0.8244	0.9619	a	-0.8155	0	0
	C=10	2	a	0.9136	0.9931	a	0.4475	0	0
		4	a	0.8994	0.9883	a	-0.1405	0	0
		6	a	0.8528	0.9778	a	-0.5483	0	0
		8	a	0.8243	0.9620	a	-0.8175	0	0

[a]Because estimation converged for all cases at this level and no trimming was done on the complete vector the $r_{\theta\hat{\theta}_c}$s and the $\bar{\theta}_c$s are the same for a given level of Nomitted across C levels

[b]Number of nonconvergent cases

Note: $\bar{\theta}_c$: ability estimates based on the complete vectors, $\bar{\theta}_o$: ability estimates based on the omission vectors

Table 3: Descriptive Statistics and Fidelity Coefficients-EAP

Method	Level	Nomitted	$r_{\theta\hat{\theta}_c}$	$r_{\theta\hat{\theta}_o}$	$r_{\hat{\theta}_c\hat{\theta}_o}$	$\bar{\hat{\theta}}_c$	$\bar{\hat{\theta}}_o$
EAP	10 Points	2	0.9183	0.9127	0.9929	0.3934	0.4721
		4	0.9097	0.8991	0.9880	-0.2457	-0.1433
		6	0.8682	0.8524	0.9773	-0.6664	-0.5711
		8	0.8328	0.8249	0.9624	-0.9428	-0.8538
	20 Points	2	0.9184	0.9129	0.9931	0.3935	0.4720
		4	0.9099	0.8992	0.9883	-0.2451	-0.1430
		6	0.8692	0.8531	0.9779	-0.6656	-0.5701
		8	0.8325	0.8245	0.9626	-0.9418	-0.8527

Note: $\bar{\theta}_C$: ability estimates based on the complete vectors, $\bar{\theta}_O$: ability estimates based on the omission vectors

Table 4: Descriptive Statistics and Fidelity Coefficients-MLE

Method	Level	Nomitted	$r_{\theta\hat{\theta}_c}$	$r_{\theta\hat{\theta}_o}$	$r_{\hat{\theta}_c\hat{\theta}_o}$	$\bar{\hat{\theta}}_c$	$\bar{\hat{\theta}}_o$	Nonconvergent[a] Compl	Nonconvergent[a] Omit
MLE	Nalt=4	2	0.9001	0.9019	0.9886	0.5395	0.4136	0	0
		4	0.8998	0.8878	0.9840	-0.2811	-0.4588	1	1
		6	0.8506	0.8208	0.9732	-0.8131	-1.0484	0	0
		8	0.8013	0.7865	0.9501	-1.1599	-1.4624	0	0
	Nalt=7	2	0.9001	0.9028	0.9880	0.5395	0.3694	0	0
		4	0.8998	0.8856	0.9822	-0.2811	-0.5318	1	2
		6	0.8506	0.8166	0.9701	-0.8131	-1.1559	0	2
		8	0.8013	0.7728	0.9433	-1.1599	-1.6268	0	0
	Ignored	2	0.9001	0.8891	0.9866	0.5395	0.6551	0	57
		4	0.8998	0.8832	0.9834	-0.2811	-0.1543	1	23
		6	0.8506	0.8311	0.9719	-0.8131	-0.7235	0	5
		8	0.8013	0.7843	0.9474	-1.1599	-1.1375	0	0
	Incorrect	2	0.9001	0.9035	0.9868	0.5395	0.3120	0	0
		4	0.8998	0.8819	0.9795	-0.2811	-0.6294	1	3
		6	0.8506	0.8095	0.9643	-0.8131	-1.3041	0	5
		8	0.8013	0.7463	0.9260	-1.1599	-1.8646	0	0

[a]Number of nonconvergent cases

Note: $\bar{\theta}_c$: ability estimates based on the complete vectors, $\bar{\theta}_o$: ability estimates based on the omission vectors

Table 5: Descriptive Statistics and Fidelity Coefficients-2PL, 3PL, NR

Nomitted	$r_{\theta\hat{\theta}_{2PL}}$	$r_{\theta\hat{\theta}_{3PL}}$	$r_{\theta\hat{\theta}_{NR}}$	$r_{\hat{\theta}_{2PL}\hat{\theta}_{NR}}$	$r_{\hat{\theta}_{3PL}\hat{\theta}_{NR}}$	$\bar{\hat{\theta}}_{2PL}$	$\bar{\hat{\theta}}_{3PL}$	$\bar{\hat{\theta}}_{NR}$	$\bar{\theta}$
2	0.9162	0.9187	0.9083	0.9904	0.9895	0.2699	0.3151	0.2794	0.3604
4	0.9095	0.9102	0.8968	0.9842	0.9831	-0.2959	-0.3143	-0.3047	-0.3335
6	0.8698	0.8693	0.8483	0.9715	0.9700	-0.6610	-0.7286	-0.6922	-0.8139
8	0.8331	0.8330	0.8086	0.9538	0.9510	-0.8968	-1.0038	-0.9499	-1.0694

Figure 1a: RMSE for Biweight ability estimation

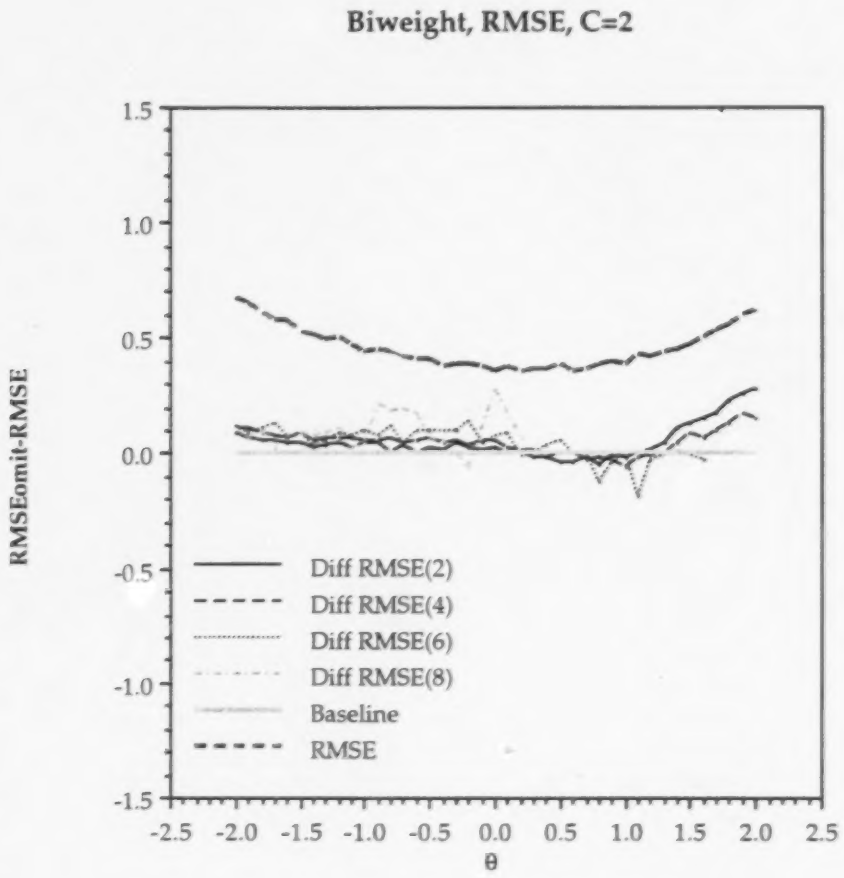


Figure 1b: RMSE for Biweight ability estimation

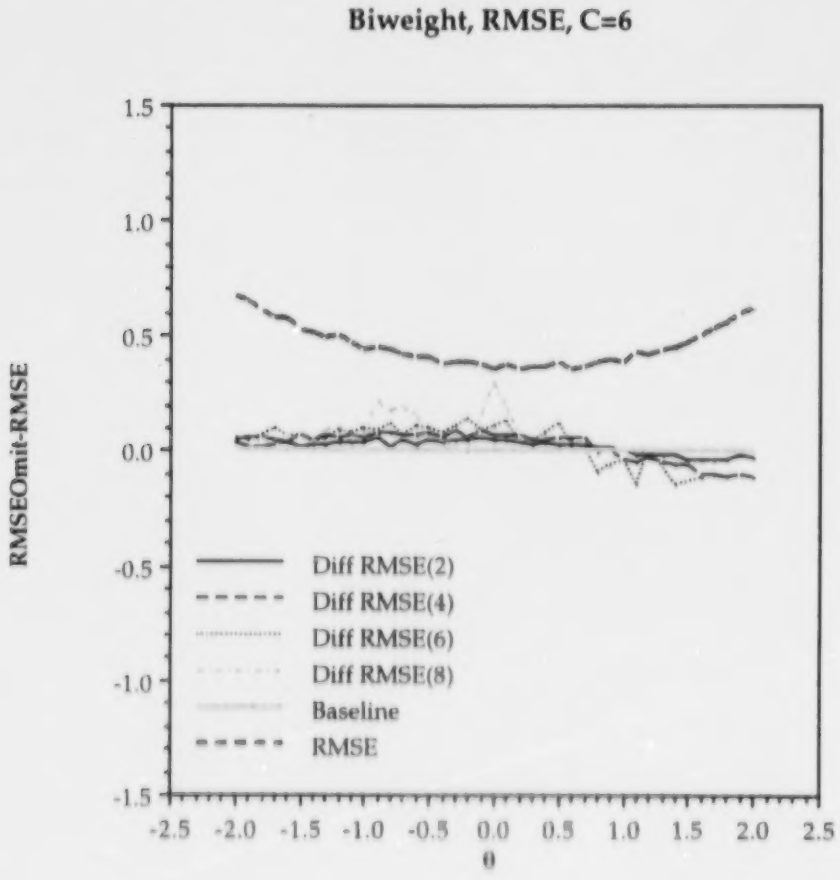


Figure 1c: RMSE for Biweight ability estimation

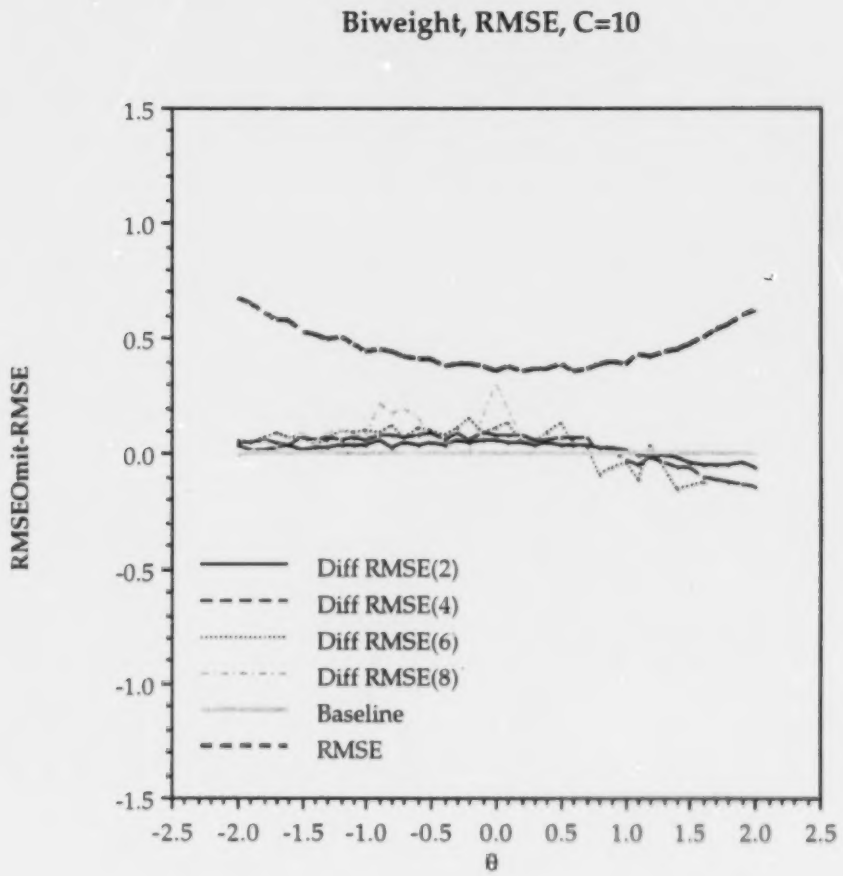


Figure 2a: Bias for Biweight ability estimation

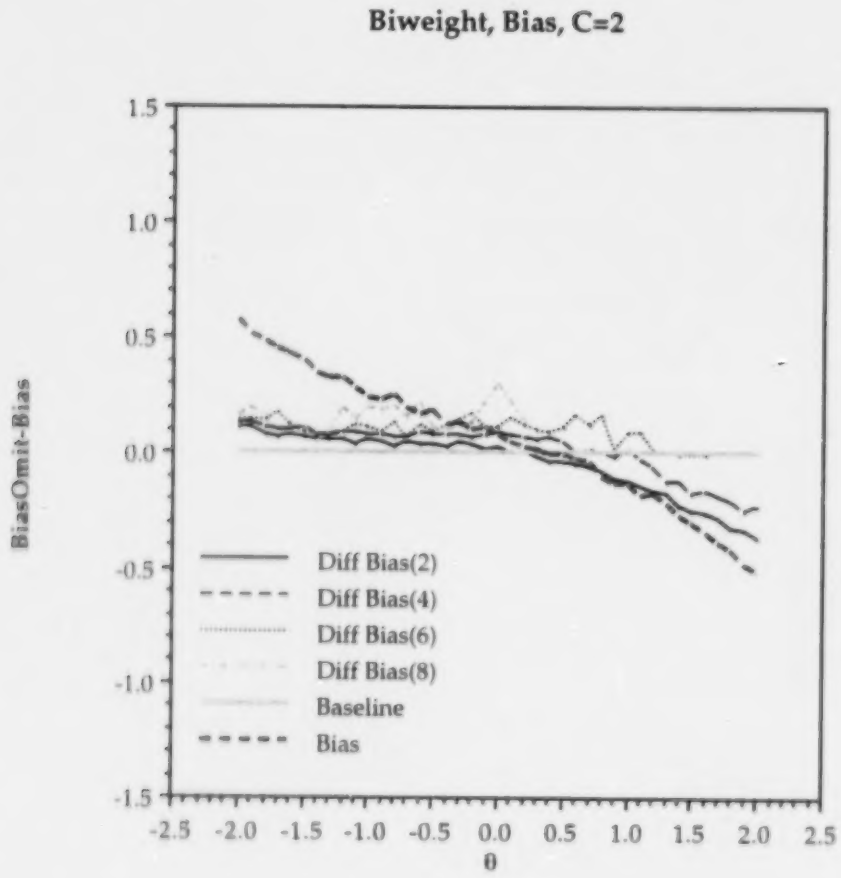


Figure 2b: Bias for Biweight ability estimation

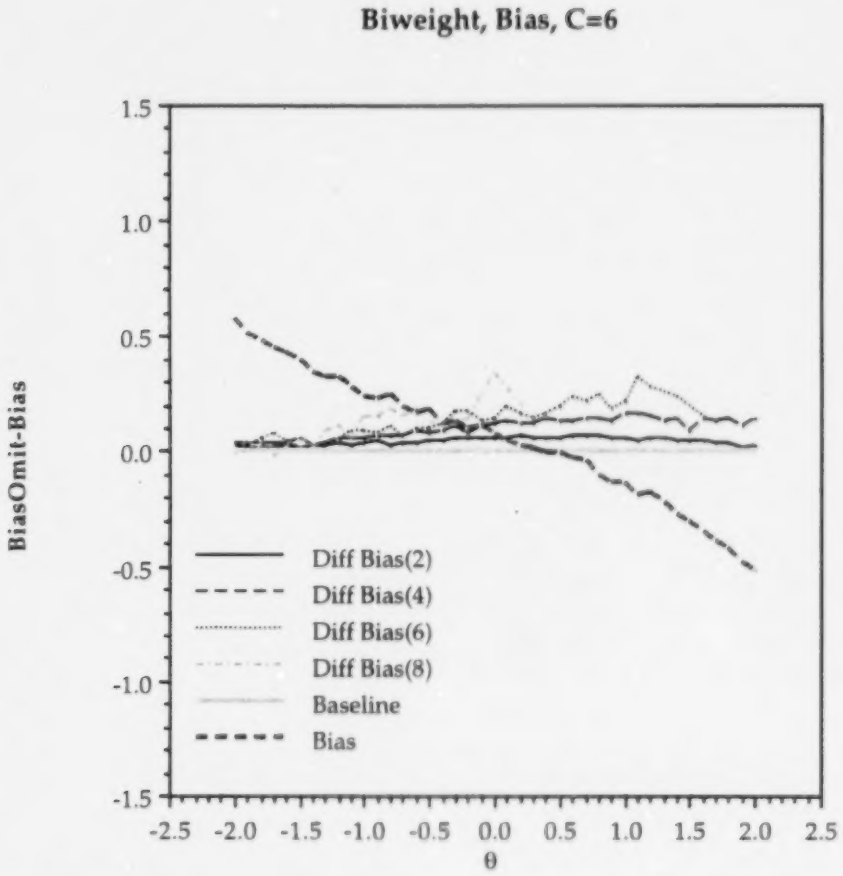


Figure 2c: Bias for Biweight ability estimation

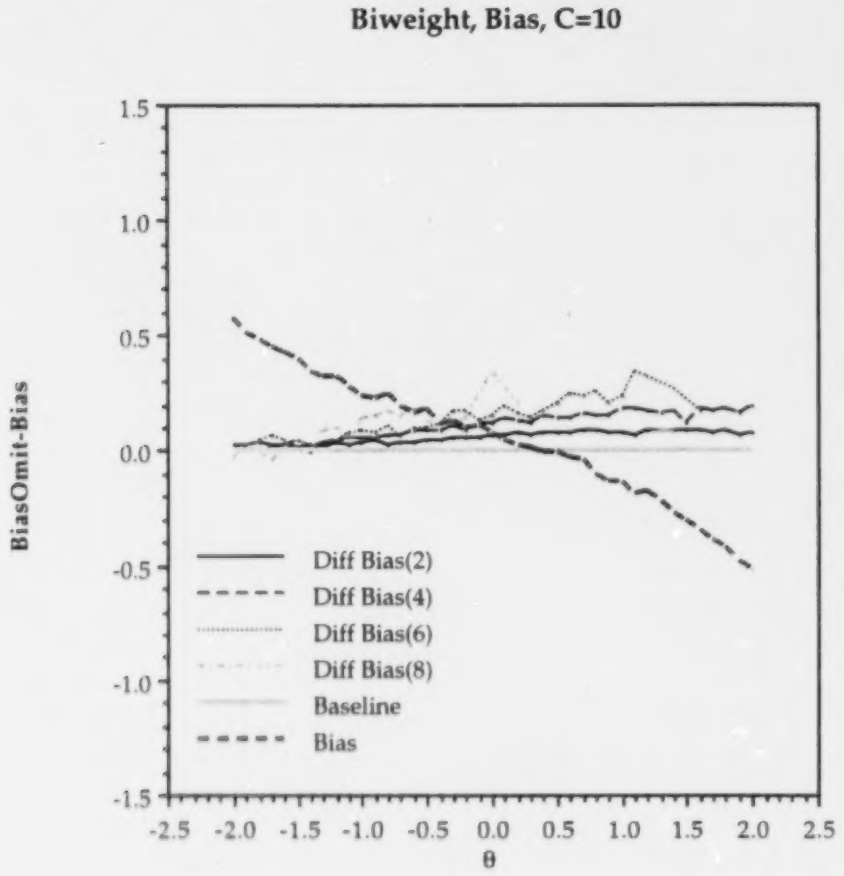


Figure 3: RMSE for EAP ability estimation

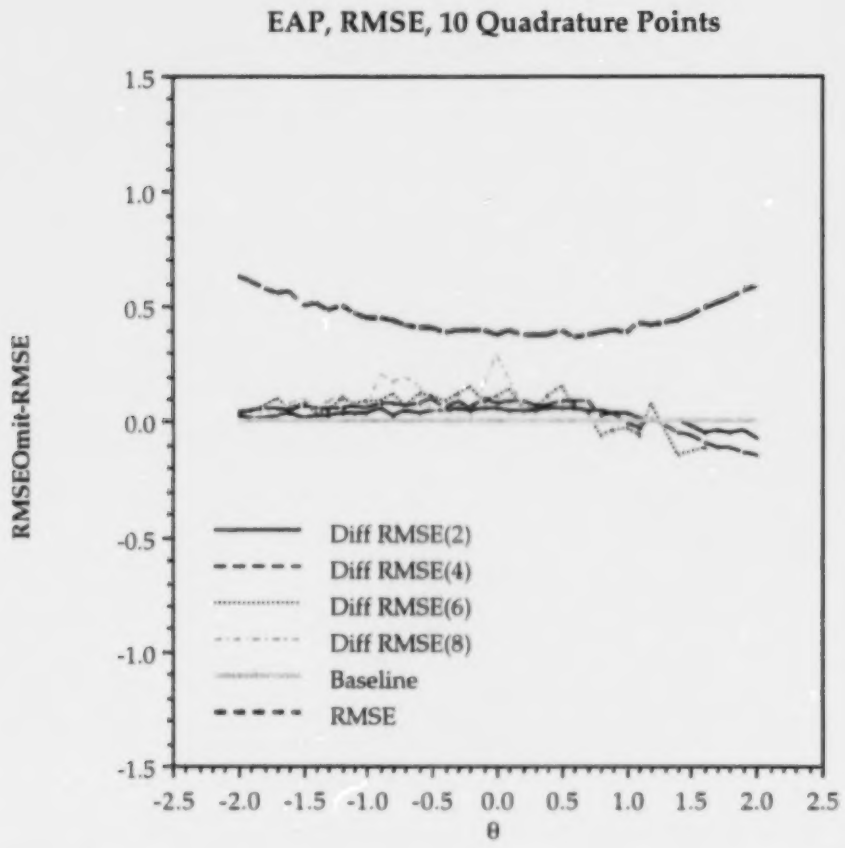


Figure 4: Bias for EAP ability estimation

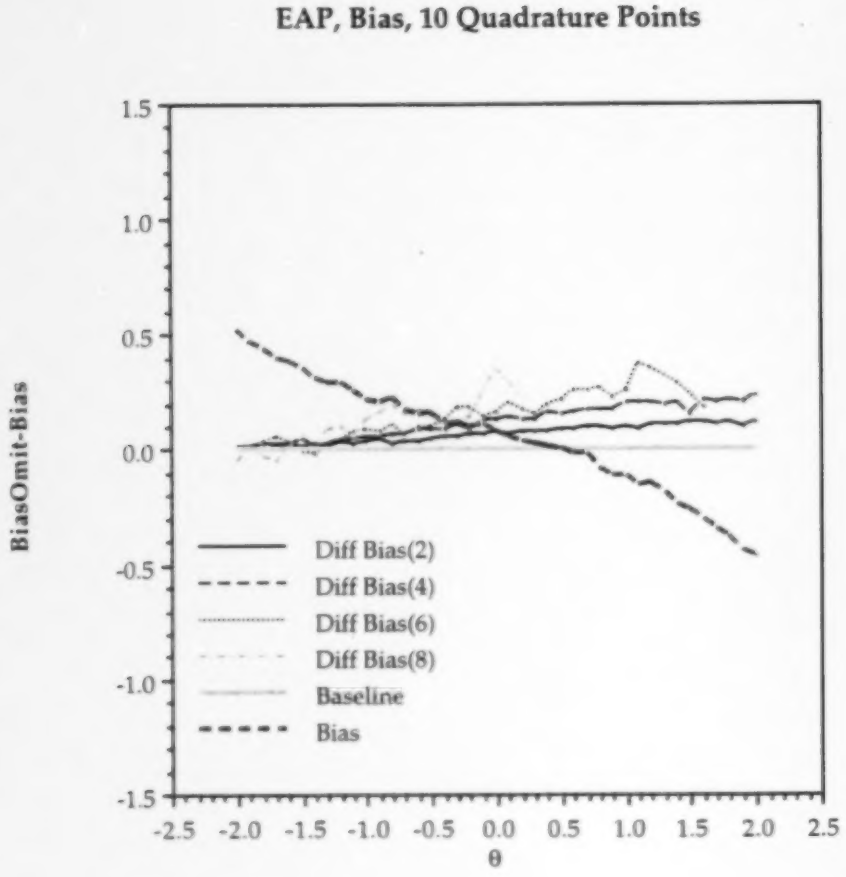


Figure 5a: RMSE for MLE ability estimation

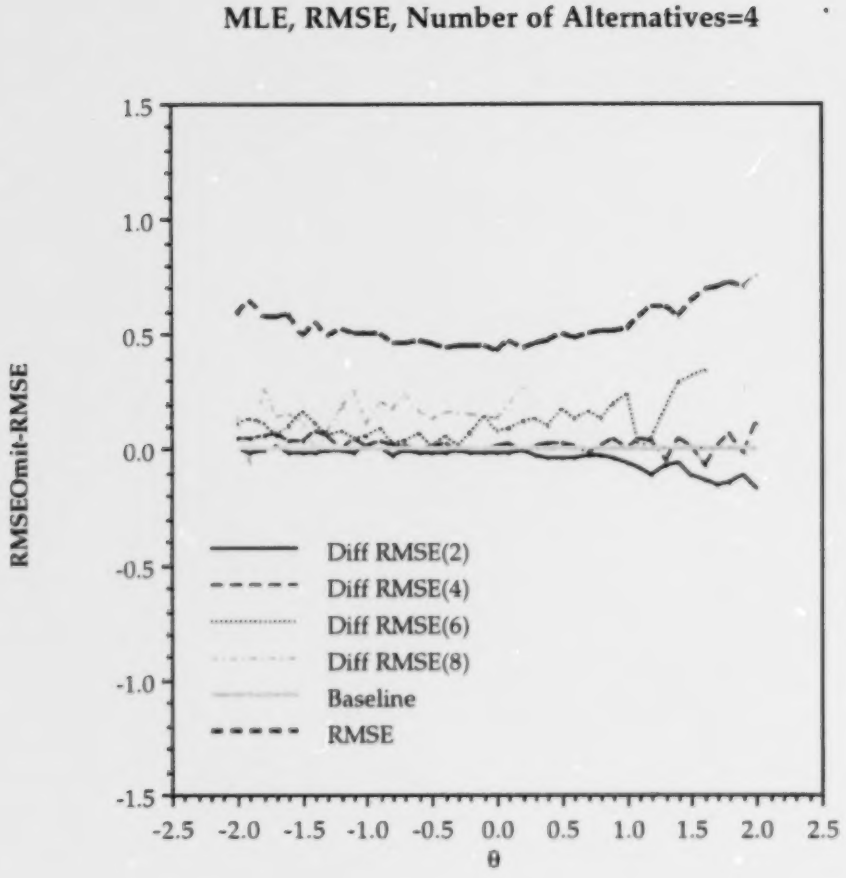


Figure 5b: RMSE for MLE ability estimation

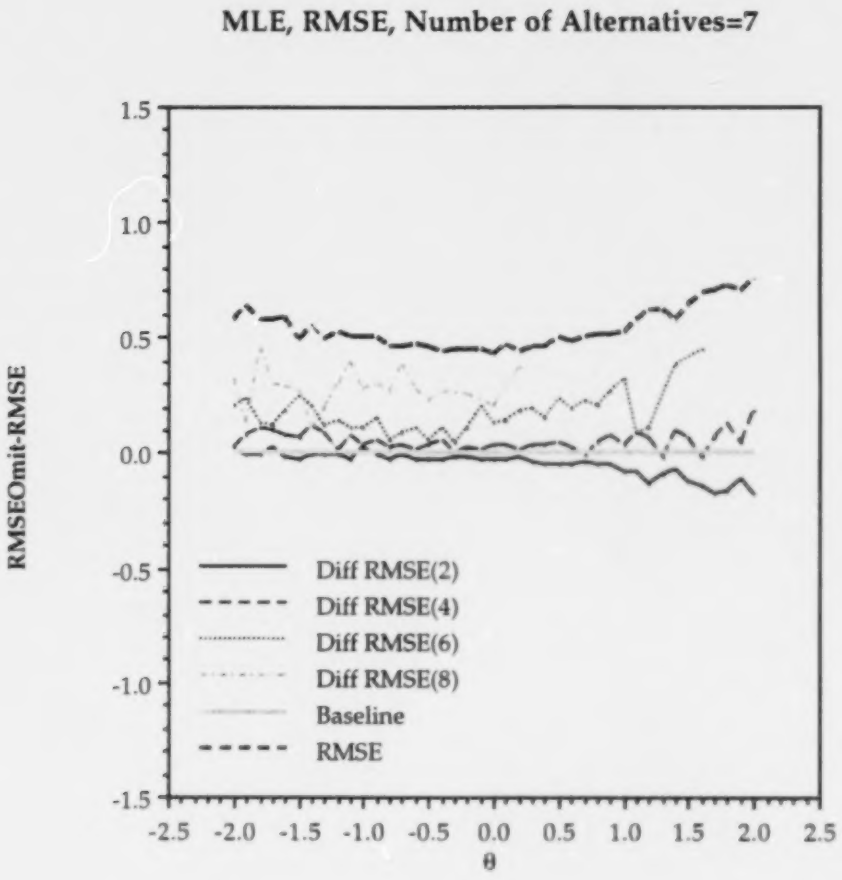


Figure 5c: RMSE for MLE ability estimation

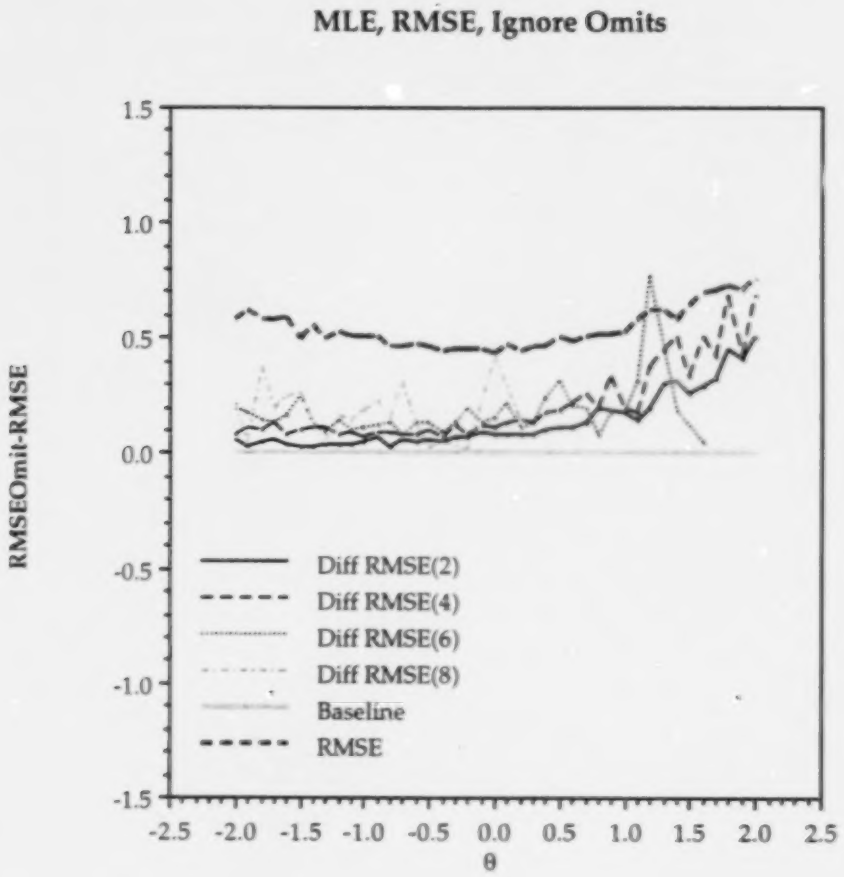


Figure 5d: RMSE for MLE ability estimation

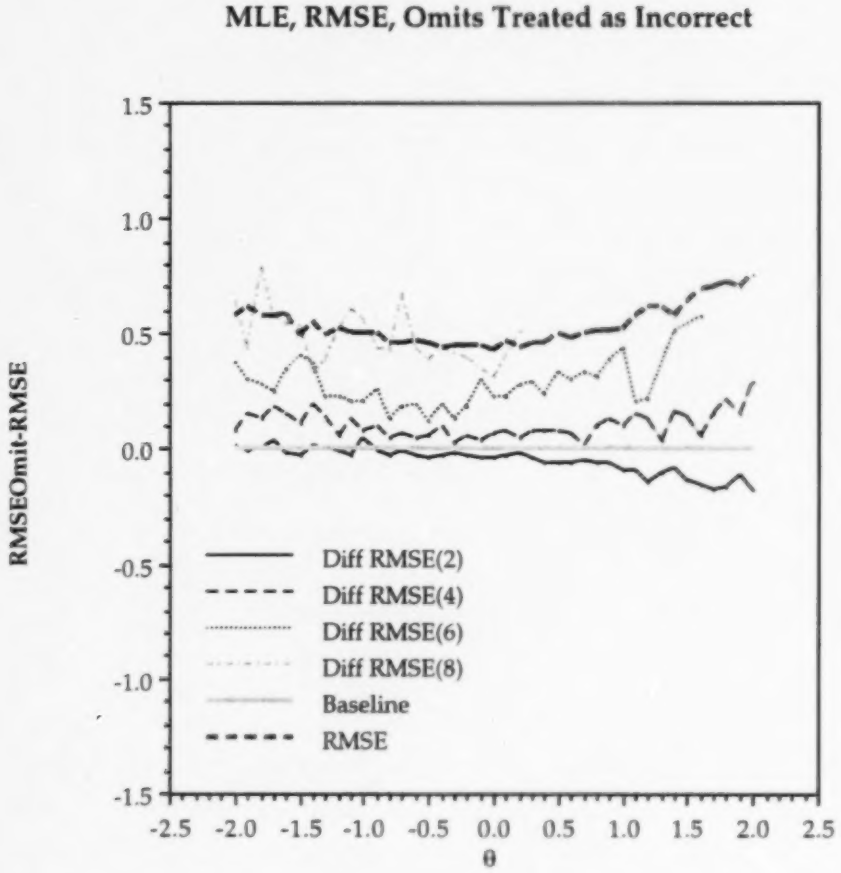


Figure 6a: Bias for MLE ability estimation

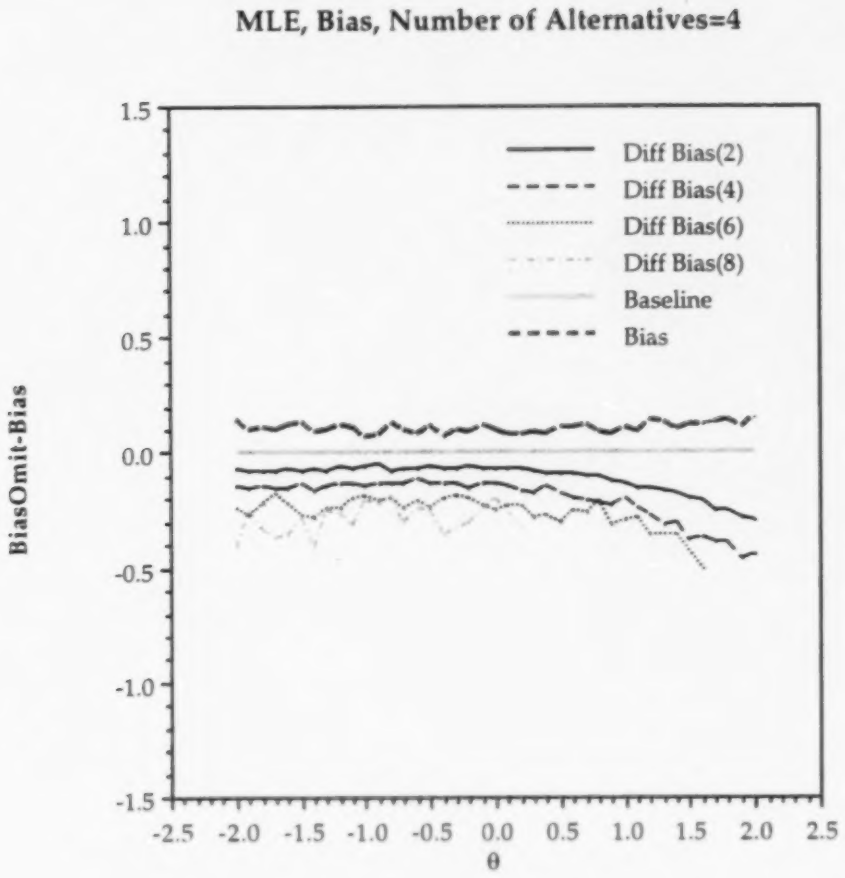


Figure 6b: Bias for MLE ability estimation

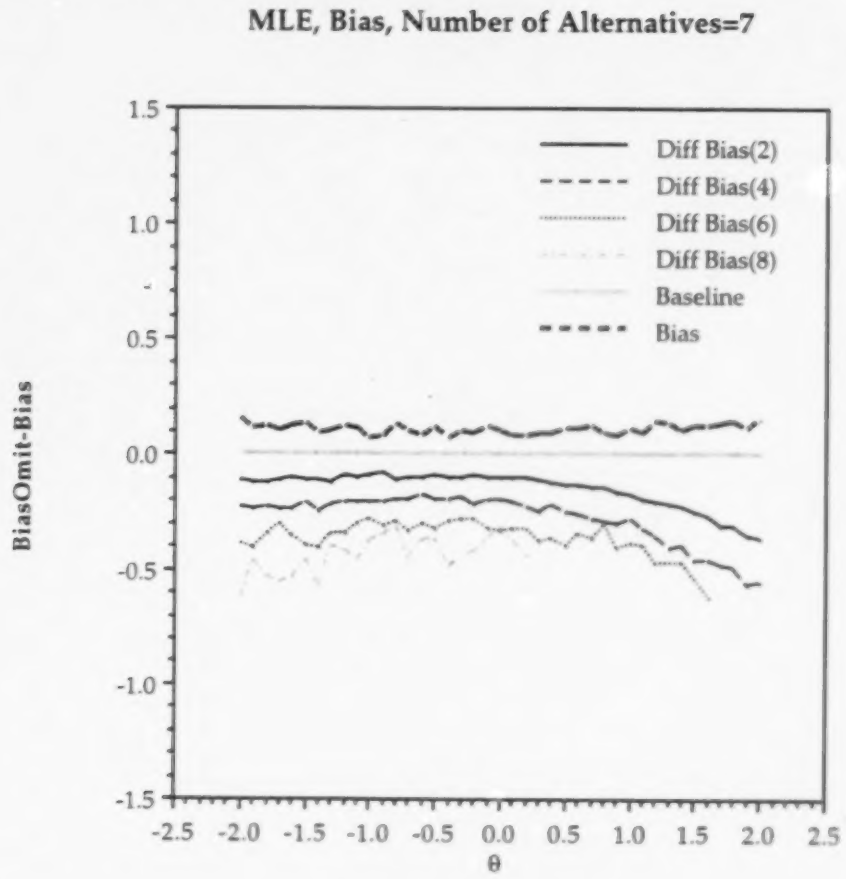


Figure 6c: Bias for MLE ability estimation

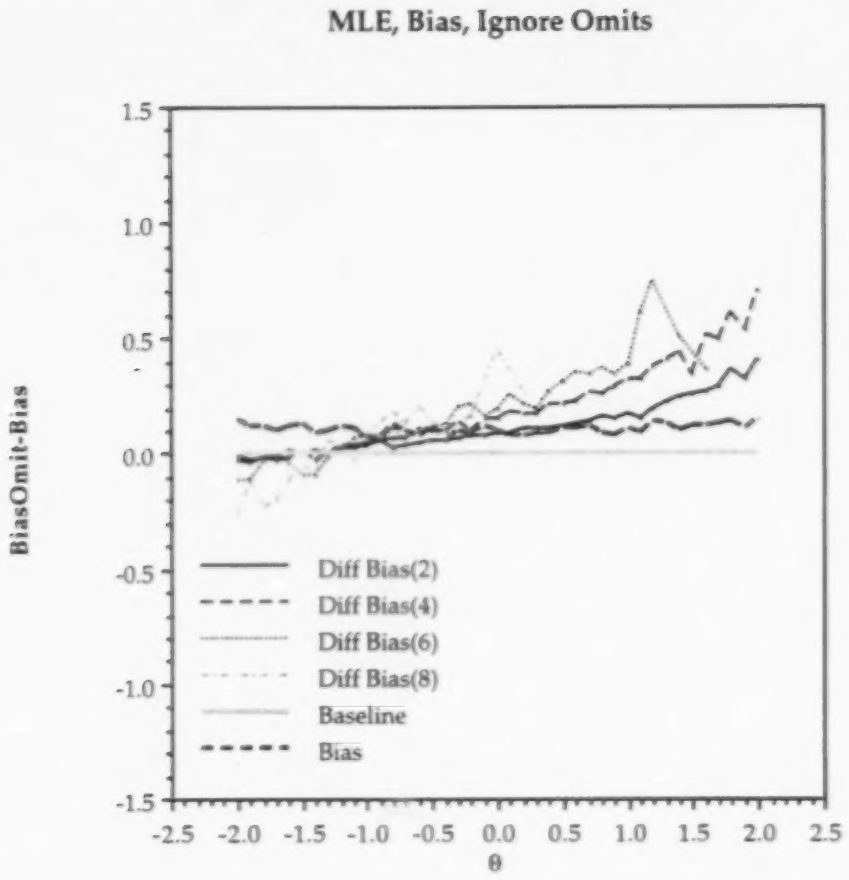


Figure 6d: Bias for MLE ability estimation

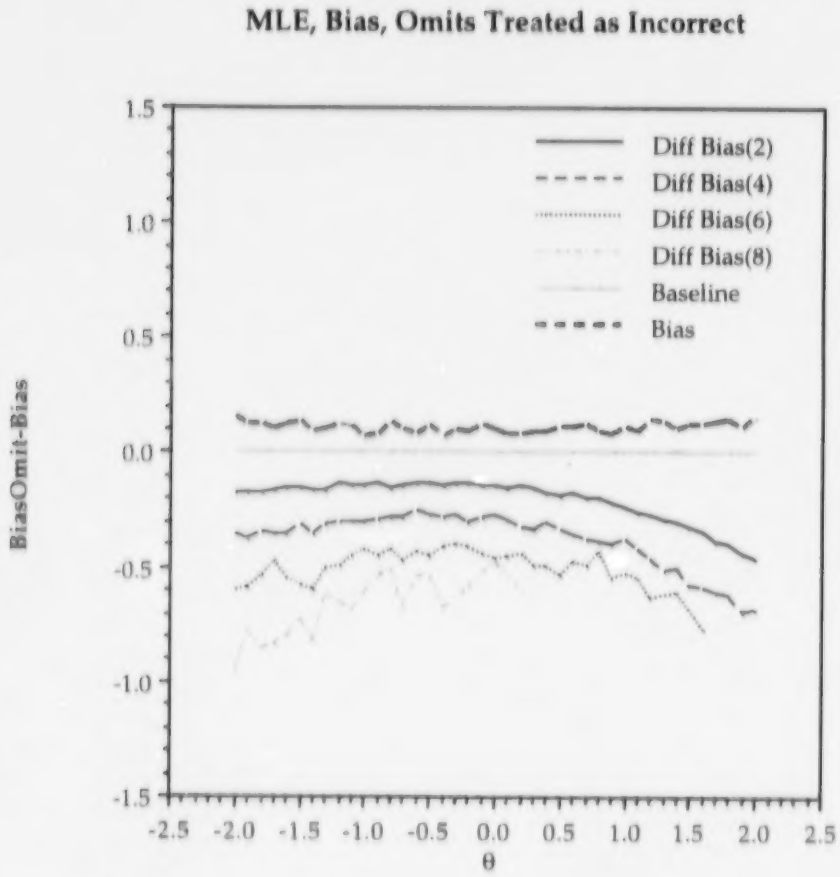


Figure 7: Average Standard Error for EAP and MLE for Complete Vectors

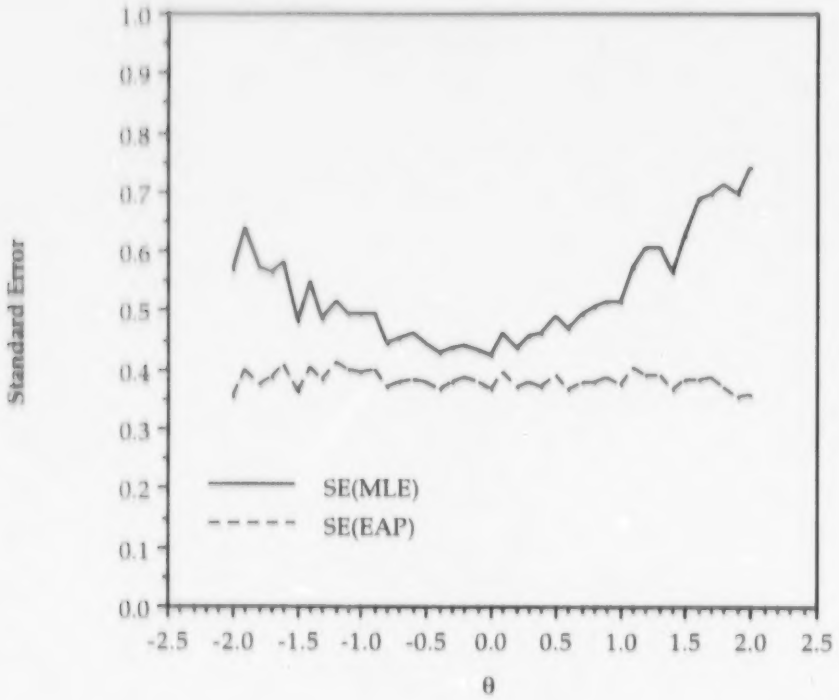


Figure 8a: RMSE for EAP ability estimation using 10 Quadrature Points and Nalt=2

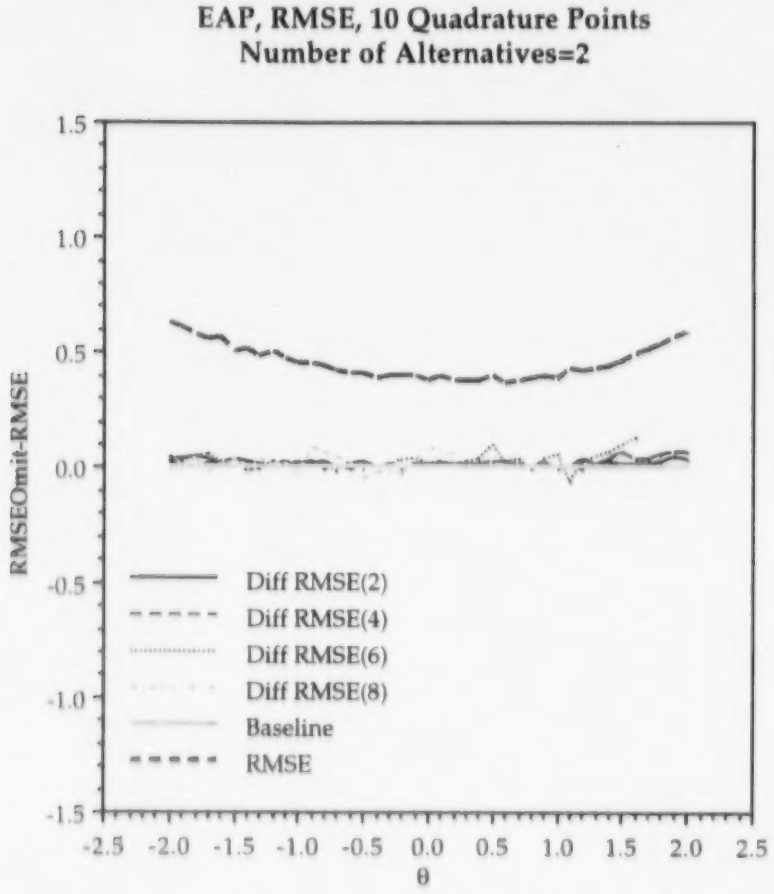


Figure 8b: Bias for EAP ability estimation using 10 Quadrature Points and the Nalt=2

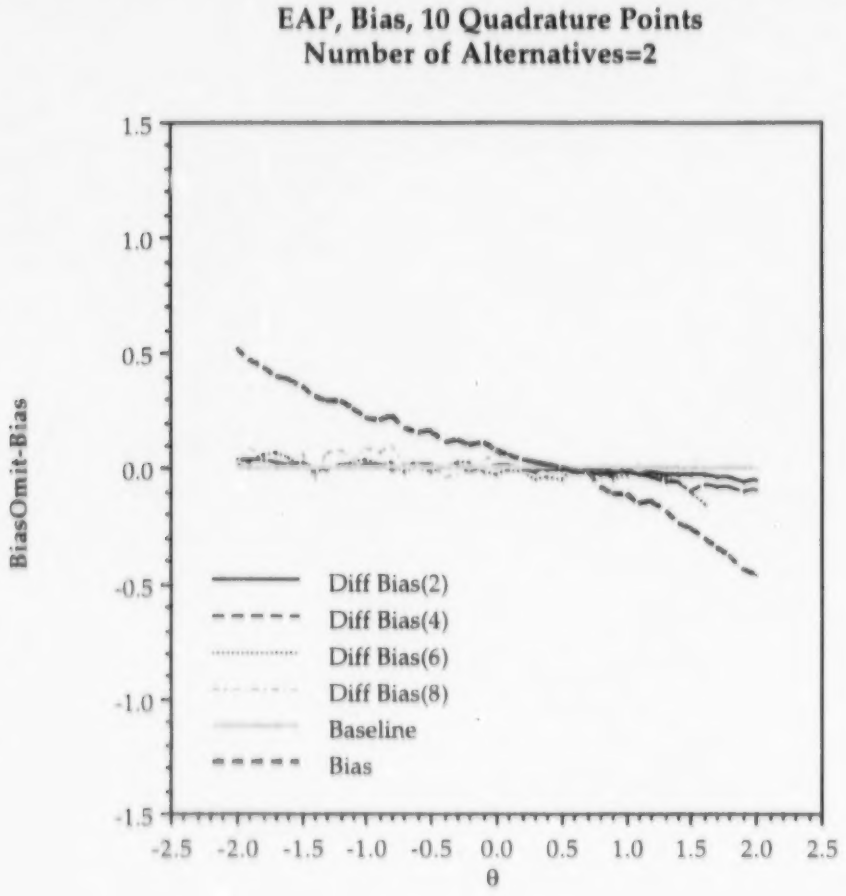


Figure 9a: RMSE for 2PL and NR models

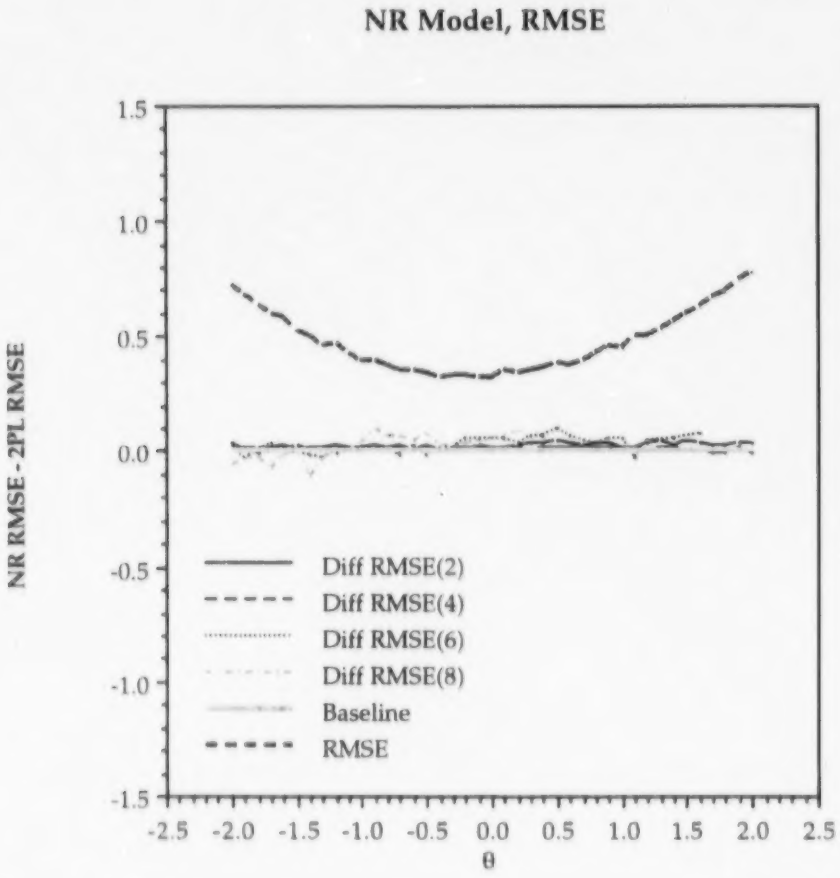


Figure 9b: Bias for 2PL and NR models

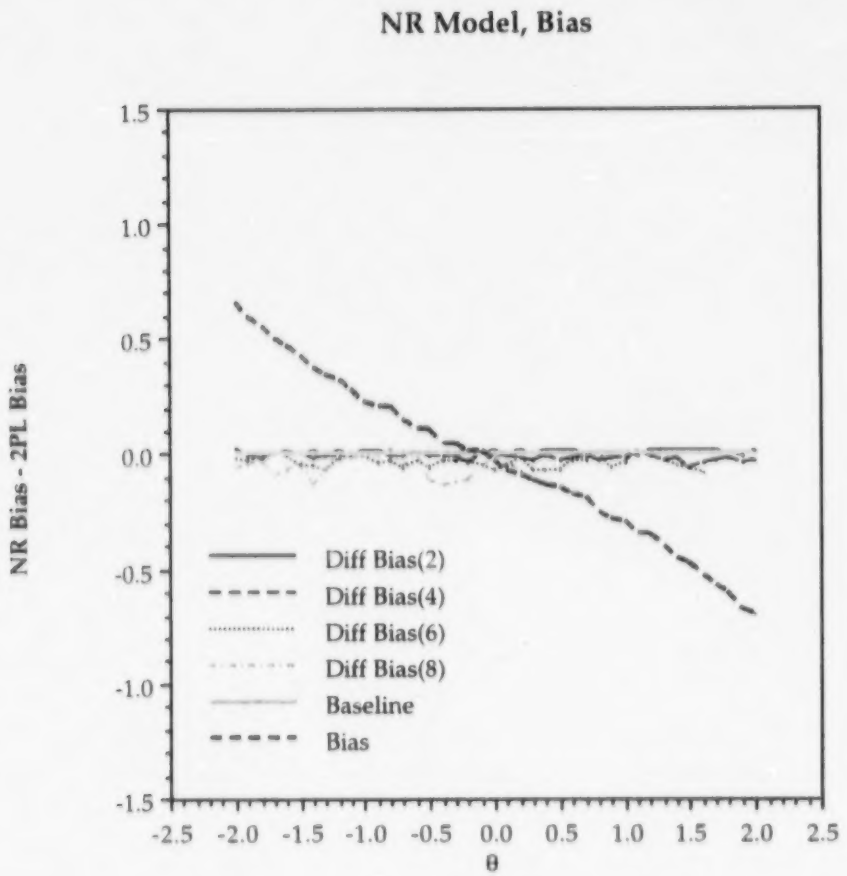


Figure 10a: RMSE for 3PL and NR models

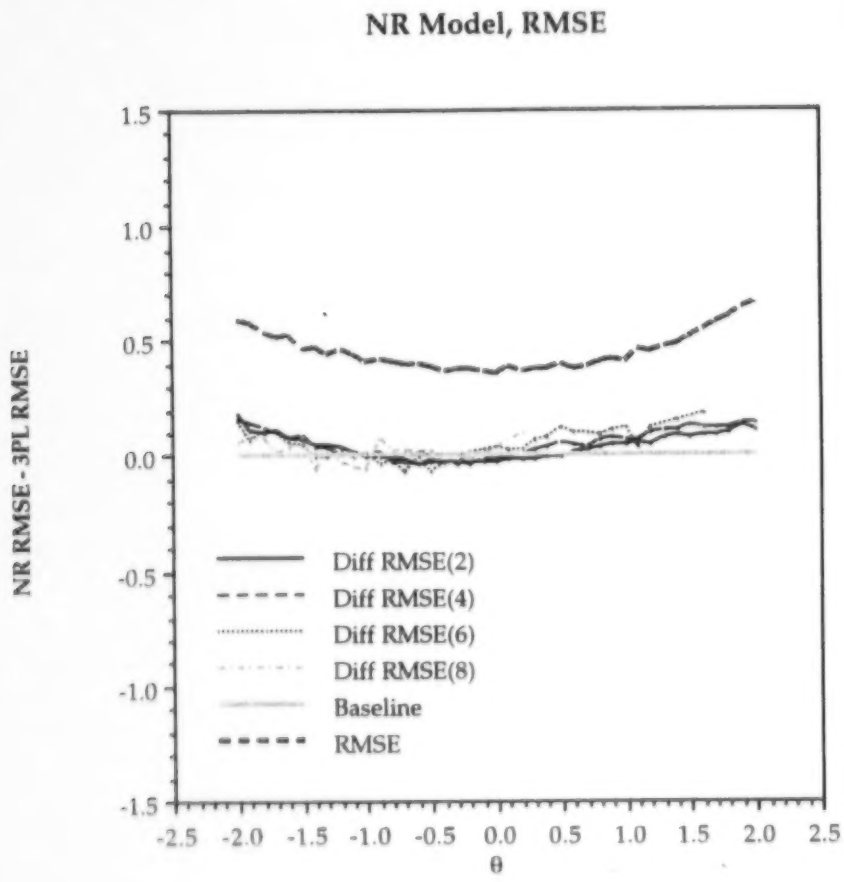
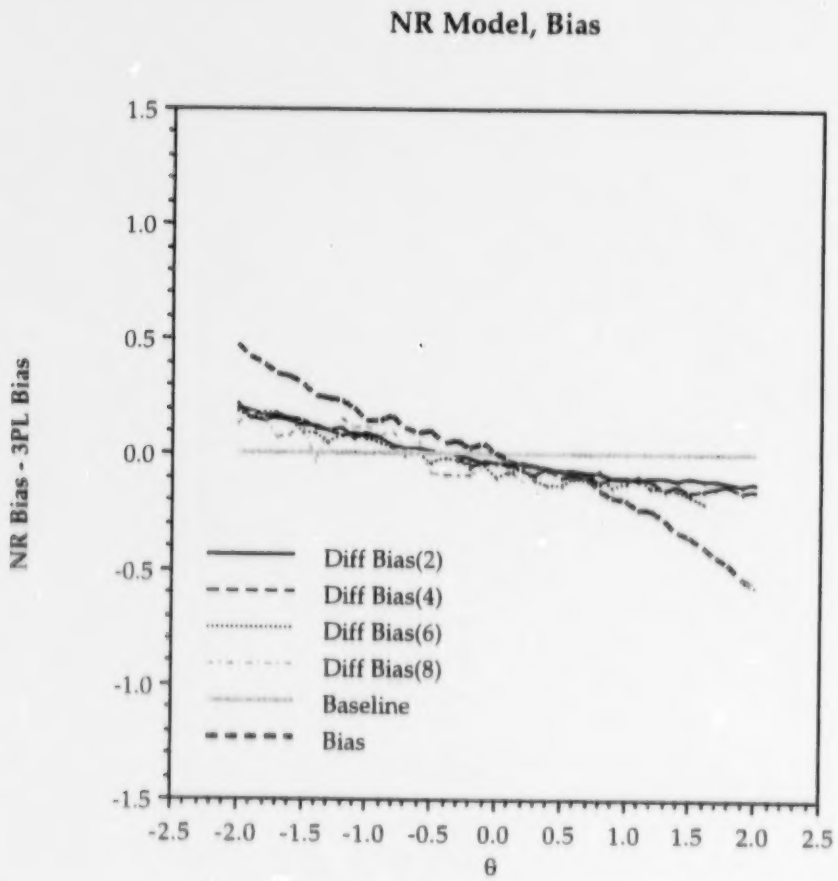


Figure 10b: Bias for 3PL and NR models



BEST COPY AVAILABLE